Bayesian Approach

This supplement describes a Bayesian approach to estimating allele frequency from RNA-seq samples. We explicitly correct for technical bias using a prior.

1 Model Specification

Let M denote *D. melanogaster* and S *D. simulans.* We propose a Bayesian model based estimation of the S allelic frequency, θ. The model combines information from each biological replicate i and corrects for possible bias attributable to technical and biological factors. Technical factors include sequencing biases, mapping bias and any other effect that casuses the measurement to be biased. Biological factors include things such as gene duplications that differ among species. The observed reads from the DNA samples will encompass both of these sources of bias, as would controls based upon mixed parental samples. The DNA control is used as the prior to correct for such biases. While we refer to DNA throughout, the same approach could be used when mixed parental RNA samples are used as a control.

In order to define the model we introduce the following notation:

$$\begin{aligned} X_i &= \text{number of reads from M in the RNA for biorep } i, \\ l_i &= \text{number of reads from S in the RNA for biorep } i, \\ Y_{i'} &= \text{number of reads from M in the DNA for biorep } i', \\ k_{i'} &= \text{number of reads from S in the DNA for biorep } i', \\ \theta &= \text{S allele frequency in the RNA,} \end{aligned}$$

$i = 1, \ldots I$, $i' = 1, \ldots I'$. I the total number of RNA biological replications and I' the number of DNA biological replications.

The following RNA model is assumed:

$$X_i | l_i, \theta \sim \text{NegativeBinomial}(l_i, \theta)$$
$$\theta | p \sim \text{Beta}((1-p)t, pt) \qquad \text{and} \qquad l_1, \ldots, l_I | \lambda \overset{\text{iid}}{\sim} \text{Poisson}(\lambda)$$
$$\lambda \sim \text{Gamma}(a_\lambda, b_\lambda)$$

The following DNA model is assumed:

$$Y_{i'}|k_{i'}, p \sim \text{NegativeBinomial}(k_{i'}, p)$$
$$p \sim \text{Beta}(v, v) \qquad \text{and} \qquad k_1, \ldots, k_{I'}|\delta \overset{\text{iid}}{\sim} \text{Poisson}(\delta)$$
$$\delta \sim \text{Gamma}(a_\delta, b_\delta)$$

The RNA and DNA models are connected by the common parameter p, the probability of sequencing and mapping the allele S in the DNA sample. In the DNA model, p is a parameter to be estimated. In contrast, in the RNA model p, is a hyperparameter.

The expectation in a heterozygous DNA sample is $p = 0.5$. When p is greater (lower) than 0.5 the bias is towards S (M). The distribution of θ in the RNA model is then centered below (above) 0.5, at $1 - p$, to correct for this.

The estimation of θ must take into account the RNA information as well as the DNA information. The RNA information enters into the RNA model directly through the sample distribution. The DNA enters into the RNA model through the hyperparameter p. Since we do not know p we are required to estimate it. In order to do so we use the DNA model. That is a posterior value of θ, θ_m will be obtained by:

1. Generate p_m from the posterior of the parameter p using the DNA model.

2. Generate θ_m from the posterior of θ in the RNA model with $p = p_m$.

The hyperparameter t plays a fundamental role. The Beta$((1 - p)t, pt)$ has mean $1 - p$ and variance $p(1 - p)/(t + 1)$, so t is a precision parameter. A standard interpretation of the hyperparameters of a beta prior, adapted to our context, is that $1 - p$ is the rate of M reads and $t + 1$ the total number (M+S) reads in an imaginary sample. If we want the prior to be as strong as the likelihood in the RNA model we require $t + 1$ to be equal to the total number of RNA reads. With such a choice, the posterior expected value of θ is [$(1 - p)$ + total #S RNA reads/total # RNA reads]/2. For example, if $p = 0.8$ and the proportion of S RNA observed reads is 0.8 as well, then the posterior expected value of of θ is 0.5. We, then, assume this value for t, i.e., $t = \sum_i (x_i + l_i) - 1$.

Here the Gamma(a, b) distribution has mean a/b. Besides, $X \sim$NegativeBinomial(l, θ), $X \in \{0, 1, \ldots, \infty\}$ represents the number of "failures" before the first l "successes" and θ the success rate. The Negative binomial distribution is a Poisson-gamma mixture. More specifically,

$$X \sim \text{NegativeBinomial}(l, \theta) \Leftrightarrow X \sim \text{Poisson}(\eta) \text{ and } \eta \sim \text{Gamma}(l, \theta/(1 - \theta)).$$

The inference of θ is based on Markov chain Monte Carlo methods. We will use this representation of the Negative Binomial distribution in the Gibbs sampler scheme that simulates from the posterior distribution of the model. That is we replace the first level of the RNA model by $X_i|\eta_i \sim \text{Poisson}(\eta_i)$ and $\eta_i \sim \text{Gamma}(l_i, \theta/(1-\theta))$. Analogously, we replace the first level of the DNA model by $Y_{i'}|\gamma_{i'} \sim \text{Poisson}(\gamma_{i'})$ and $\gamma_{i'} \sim \text{Gamma}(k_{i'}, p/(1-p))$.

Other notes:

1. The prior for p in the DNA model, beta(v, v), has mean $1/2$ and variance $1/(4(v+1))$. The parameter v is free to choose. We choose $v = 1$ so the prior is uniform.

2. λ (δ) is the expected number of RNA (DNA) S reads for each one of the biological replications. We recommend a vague prior for the gamma distribution of the parameters λ and δ. We used $a_\lambda = a_\delta = 1/2$ and $b_\lambda = b_\delta = 1/2$. The model is not sensitive to the value of these hyperparameters (not shown).

2 Complete Conditionals for Gibbs Sample

The likelihood functions of both models are:

$$
\begin{aligned}
& L_{RNA}(\theta, \eta, \lambda, \eta | p, x, l, k) \\
& \quad \propto \quad \left\{ \prod_i \text{Poisson}(x_i | \eta_i) \right\} \left\{ \prod_i \text{Gamma}(\eta_i | l_i, \theta/(1-\theta)) \right\} \\
& \quad \times \quad \left\{ \prod_i \text{Poisson}(l_i | \lambda) \right\} \{ \text{Gamma}(\lambda | a_\lambda, b_\lambda) \} \\
& \quad \times \quad \{ \text{Beta}(\theta | (1-p)t, pt) \} \\
& \quad \propto \quad \left\{ \prod_i \frac{e^{-\eta_i} \eta_i^{x_i}}{x_i!} \right\} \left\{ \prod_i \frac{\left(\frac{\theta}{1-\theta}\right)^{l_i}}{(l_i - 1)!} \eta_i^{l_i - 1} \exp\left(-\frac{\theta}{1-\theta} \eta_i \right) \right\} \\
& \quad \times \quad \left\{ \prod_i \frac{e^{-\lambda} \lambda^{l_i}}{l_i!} \right\} \{ \lambda^{a_\lambda - 1} e^{-b_\lambda \lambda} \} \\
& \quad \times \quad \left\{ \theta^{(1-p)t-1} (1-\theta)^{pt-1} \right\}
\end{aligned}
$$

and

$$
\begin{aligned}
& L_{DNA}(p, \gamma, \delta | y, k) \\
& \quad \propto \quad \left\{ \prod_{i'} \text{Poisson}(y_{i'} | \gamma_{i'}) \right\} \left\{ \prod_{i'} \text{Gamma}(\gamma_{i'} | k_{i'}, p/(1-p)) \right\} \\
& \quad \times \quad \left\{ \prod_{i'} \text{Poisson}(k_{i'} | \delta) \right\} \{ \text{Gamma}(\delta | a_\delta, b_\delta) \} \\
& \quad \times \quad \text{Beta}(p | v, v) \\
& \quad \propto \quad \left\{ \prod_{i'} \frac{e^{-\gamma_{i'}} \gamma_{i'}^{y_{i'}}}{y_{i'}!} \right\} \left\{ \prod_{i'} \frac{\left(\frac{p}{1-p}\right)^{k_{i'}}}{(k_{i'} - 1)!} \gamma_{i'}^{k_{i'} - 1} \exp\left(-\frac{p}{1-p} \gamma_{i'} \right) \right\} \\
& \quad \times \quad \left\{ \prod_{i'} \frac{e^{-\delta} \delta^{k_{i'}}}{k_{i'}!} \right\} \{ \delta^{a_\delta - 1} e^{-b_\delta \delta} \} \\
& \quad \times \quad \{ p^{v-1} (1-p)^{v-1} \}
\end{aligned}
$$

2.1 Complete Conditionals for RNA Model

For the RNA model we get the following full conditionals for the Gibbs sampler:

$$\begin{aligned} p(\eta_i|all) &\propto e^{-\eta_i}\eta_i^{x_i}\eta_i^{l_i-1}\exp\left(-\frac{\theta}{1-\theta}\eta_i\right) \\ &\propto \eta_i^{x_i+l_i-1}\exp\left(-\left[1+\frac{\theta}{1-\theta}\right]\eta_i\right) \\ &\propto \text{Gamma}(\eta_i|x_i+l_i, 1/(1-\theta)) \end{aligned}$$

$$\begin{aligned} p(\lambda|all) &\propto \left\{\prod_i e^{-\lambda}\lambda^{l_i}\right\}\lambda^{a_\lambda-1}e^{-b_\lambda\lambda} \\ &\propto e^{-(b_\lambda+I)\lambda}\lambda^{a_\lambda+l_\cdot-1} \\ &\propto \text{Gamma}(\lambda|a_\lambda+l_\cdot, b_\lambda+I) \end{aligned}$$

$$\begin{aligned} p(\theta|all) &\propto \left\{\prod_i\left(\frac{\theta}{1-\theta}\right)^{l_i}\exp\left(-\frac{\theta}{1-\theta}\eta_i\right)\right\}\times\theta^{(1-p)t-1}(1-\theta)^{pt-1} \\ &\propto \left(\frac{\theta}{1-\theta}\right)^{l_\cdot}\exp\left(-\frac{\theta}{1-\theta}\eta_\cdot\right)\times\theta^{(1-p)t-1}(1-\theta)^{pt-1} \\ &\propto \left(\frac{\theta}{1-\theta}\right)^{l_\cdot+(1-p)t-1}(1-\theta)^{pt-1+(1-p)t-1}\exp\left(-\frac{\theta}{1-\theta}\eta_\cdot\right) \\ &\propto \left(\frac{\theta}{1-\theta}\right)^{l_\cdot+(1-p)t-1}(1-\theta)^{t-2}\exp\left(-\frac{\theta}{1-\theta}\eta_\cdot\right) \end{aligned}$$

Setting $u=\theta/(1-\theta)$ we get that $\theta=u/(1+u)=1-1/(1+u)=1-(1+u)^{-1}$ and $d\theta=(1+u)^{-2}\,du$ and

$$\begin{aligned} p(u|all) &\propto u^{l_\cdot+(1-p)t-1}\left(1-\frac{u}{1+u}\right)^{t-2}\exp\left(-u\eta_\cdot\right)(1+u)^{-2} \\ &\propto u^{l_\cdot+(1-p)t-1}\left(\frac{1}{1+u}\right)^{t-2}(1+u)^{-2}\exp\left(-u\eta_\cdot\right) \\ &\propto u^{l_\cdot+(1-p)t-1}(1+u)^{2-t}(1+u)^{-2}\exp\left(-u\eta_\cdot\right) \\ &\propto \frac{1}{(1+u)^t}u^{l_\cdot+(1-p)t-1}\exp\left(-u\eta_\cdot\right) \\ &\propto \frac{1}{(1+u)^t}\text{Gamma}(u|l_\cdot+(1-p)t, \eta_\cdot) \end{aligned}$$

A Metropolis-Hasting algorithm was implemented to simulate from the posterior distribution of θ: Let θ_i^m the value of θ_i in the m^{th} Gibbs iteration:

1. Set $u = \theta^m/(1-\theta^m)$

2. Generate $u^c \sim \text{Gamma}(l. + (1-p)t, \eta.)$

3. Compute

$$\alpha = \frac{p(u^c|all)}{p(u|all)} \times \frac{\text{Gamma}(u|l. + (1-p)t, \eta.)}{\text{Gamma}(u^c|l. + (1-p)t, \eta.)} = \left(\frac{1+u}{1+u^c}\right)^t$$

4. Generate $w \sim uniform(0,1)$

5. If $w < \min\{\alpha, 1\}$ set θ^{m+1} equal to $u^c/(1+u^c)$ otherwise set $\theta^{m+1} = \theta^m$.

Note that a disadvantage of the MH algorithm described above is the lack of a tuning parameter to control the acceptance ratio. If t is large, as it will be the case in our application, the algorithm above is not good. The acceptance ratio is too low. We then appeal to the standard MH. That is we replace steps 2 and 3 by:

2. Generate $u^c \sim \text{Normal}(u, \sigma_\theta^2)$ The standard deviation of the "proposed distribution", σ_θ, is chosen such that the acceptance ratio is between 0.25 and 0.75.

3. If u^c is less than zero set $\alpha = 0$ otherwise,

$$\alpha = \frac{p(u^c|all)}{p(u|all)} = \left(\frac{1+u}{1+u^c}\right)^t \frac{\text{Gamma}(u^c|l. + (1-p)t, \eta.)}{\text{Gamma}(u|l. + (1-p)t, \eta.)}$$

2.2 Complete Conditionals for DNA Model

For the DNA model we get the following full conditionals for the the Gibbs sampler: Analogous to η_i in the RNA model,

$$p(\gamma_{i'}|all) \propto \text{Gamma}(\gamma_{i'}|y_{i'} + k_{i'}, 1/(1-p))$$

with $k. = \sum_{i'} k_{i'}$

Analogous to λ,

$$p(\delta|all) \propto \text{Gamma}(\delta|a_\delta + k., b_\delta + I')$$

with $l. = \sum_i l_i$, and finally,

$$\begin{aligned} p(p|all) &\propto \left\{\prod_{i'} \left(\frac{p}{1-p}\right)^{k_{i'}} \exp\left(-\frac{p}{1-p}\gamma_{i'}\right)\right\} \times p^{v-1}(1-p)^{v-1} \\ &\propto \left\{\frac{p}{1-p}\right\}^{k.} p^{v-1}(1-p)^{v-1} \exp\left(-\frac{p}{1-p}\gamma.\right) \end{aligned}$$

Again we use a standard Metropolis-Hasting algorithm to simulate from the complete conditional posterior distribution of p. Let p^m the value of p in the m^{th} Gibbs iteration:

1. Sample $p^c \sim N(p, \sigma_p^2)$ where σ_p^2 is chosen by trial and error such that the acceptance ratio of this MH algorithm is between 0.25 and 0.75.

2. If $p^c < 0$ set $\alpha = 0$ otherwise, $\alpha = p(p^c|all)/(p|all)$

3. Generate $w \sim \text{Uniform}(0, 1)$ and set $p^{m+1} = p^c$ if $w < \alpha$, otherwise, set $p^{m+1} = p^m$.

3 Experimental Analysis

Experimental data were analyzed using the above Bayesian model. Exons were filtered according to criteria described in the main text (n=14,452) and included a coverage floor of 100 reads per exon. As expected exonic coverage was greater for RNA than DNA and there were more exonic regions with more reads in the RNA (Figure S3).

There is no obvious bias in the estimate of θ across the range of read counts in these data (Figure S4). As expected from the model specifications, the width of the 95% credible interval (CI) decreases as the number reads for RNA/DNA increases (Figure S5). For DNA, once the number of reads is above $\sim 3,500$, the average CI width is 0.034. This average exon size for these exons is $3,611$ nt, for coverage of greater than 52 reads per nucleotide (Figure S4).

In the RNA, the average credible interval width declines as the number of RNA reads increases until approximately $\sim 10,000$. When RNA coverage is very high, there is an increased variation of the CI width and some intervals are quite large. This is likely a result of the mean variance relationship in the data, and the low number of independent samples.

The results presented in the main text are for $1,000$ samples from the posterior distribution. Choosing $1,000$ samples from the posterior is common. Yet, if the posterior distributions are wide this may result in lack of precision in determining the width of the 95% credible interval. This did not appear to be a general problem for these data (Figure S5). Seven outliers were identified in the RNA (Figure S5) which are likely due to lack of precision in the estimation of the posterior. To determine whether $1,000$ samples from the posterior was sufficient, the Gibbs sampler was rerun using $10,000$ samples from the posterior(Figure S5). As expected, the width of the credible interval for the seven outlying points was better estimated, but the majority of the credible intervals were unchanged. Importantly, the inference for these 7 points remained unchanged. We also examined all $14,452$ genes for changes in inference between $10,000$ and $1,000$ samples. While there were some changes for estimates close to 0.5, the number of genes significantly different from 0.5 remained almost identical, and the bias (mel or sim) remained unchanged. Indicating that $1,000$ samples from the posterior were sufficient to accurately estimate the 95% credible interval in these data. Running the Gibbs sampler for $1,000$ estimates of the posterior of each gene required about one days worth of computation. The time increase is linear with the number of samples. As with any Bayesian approach, the number of samples from the posterior is important and care should be taken to examine the behavior of the estimates and credible intervals to insure that enough samples from the posterior have been taken.

4 Simulation Study

We simulated data based on the experiment presented in this paper. An arbitrary set of 200 exons whose three S RNA and DNA counts were positive (i.e. $l_i, k_{i'} > 0,$ for $i, i' = 1, 2, 3$) were selected. Based on these counts, for every exon we simulated three RNA samples from M, this is,

$$x_i^s \sim \text{NegativeBinomial}(l_i, \tau) \quad \text{for } i = 1, 2, 3.$$

In terms of the model, an exon exhibits AI if $\tau \neq p$. We set the simulation truth value of p equal to the proportion of reads from S in the DNA sample, i.e., $p = \sum_{i'} k_{i'} / \sum_{i'} (k_{i'} + y_{i'})$. Based on p and the observed S DNA reads, we simulated three M DNA samples,

$$y_{i'}^s \sim \text{NegativeBinomial}(k_{i'}, p) \quad \text{for } i' = 1, 2, 3.$$

In the simulated data set the first 100 exons have allelic imbalance. We set $\tau = 0.5 + 0.2 * \cos(4\pi(p - 1))$. The other 100 exons do not exhibit allelic imbalance, in other words we set $\tau = p$. The simulated data are then the same as the real data except for the number of M reads (the $x_i s$ and $y_{i'} s$). For the simulated data we applied the proposed procedure and computed the central 95% credible interval (CI) for θ. If this CI does not contain 0.5 we decide the exon has in AI. Figure S6 shows these CIs. There are 17 false negatives and 2 false positive. The power to detect AI is proportional to the width of the CI and depends on the number RNA reads. Therefore, it is not surprising, we are not able to detect AI when the RNA coverage is low. We also notice that 6 out of the 100 exons with AI, exhibit prominent AI (the CI for θs are completely contained in either (0,0.4) or (0.6,1)). While, in the real data set analyzed in this paper approximately 3% of the exons flagged with AI, exhibit extreme AI.

Figure S8 shows estimated posterior probabilities of θ under four scenarios for observed (not simulated) exons. These scenarios represent different levels of the total number of RNA and DNA reads. Figure S8 shows with fewer DNA and RNA reads, there is a flattening of the posterior distribution. We now explore the estimates of θ with and without AI. Let $\tilde{\theta}$ be the posterior expected value of θ. Figure S7 depicts the density of $\theta_{\max} = \max\{\tilde{\theta}, 1 - \tilde{\theta}\}$, i.e., a smoothed version of the histogram of the estimates of $\theta_{\max}$. We consider this $\theta_{\max}$ since values of this quantity greater than 1/2 indicate AI (regardless of the direction). The left panel of Figure S7 depicts the estimated density of $\theta_{\max}$ under AI. The right panel is the density of $\theta_{\max}$ corresponding to the exons without AI. As expected, in contrast to the $\theta_{\max}$s of exons without AI, the $\theta_{\max}$s corresponding to exons with AI tend to be greater than 1/2.

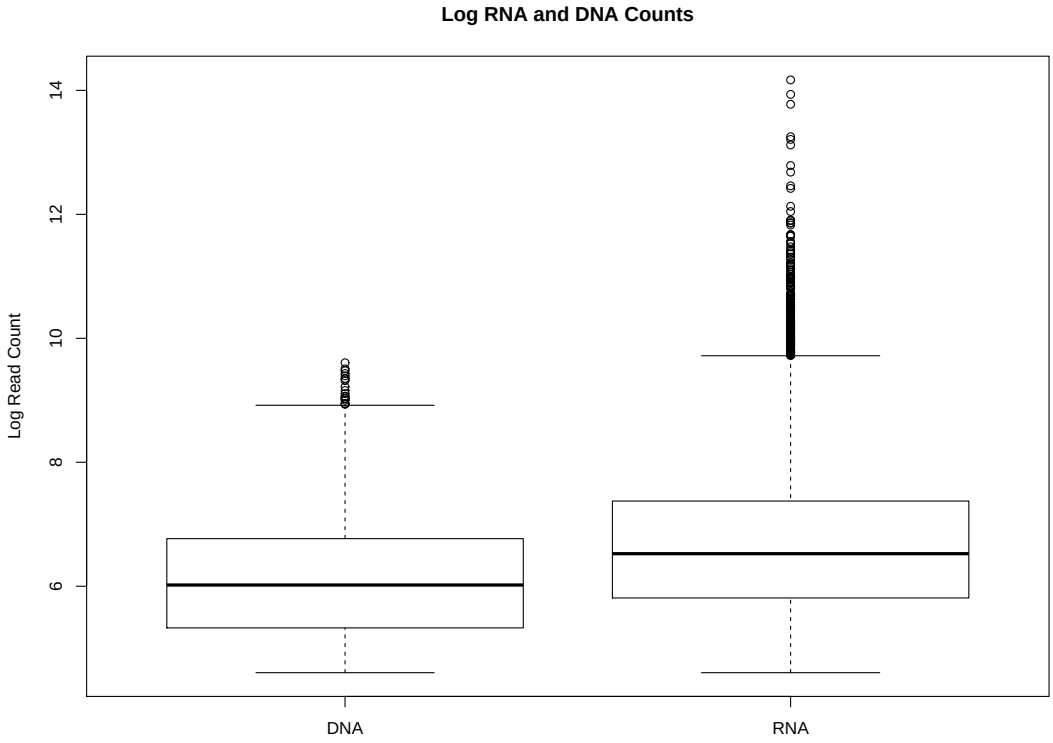


Figure S3: Distribution of the number of reads from analyzed exons (n=14,452). Counts are summed over all observations and the natural log of the sum was taken. Median number of DNA (RNA) reads 412 (682).

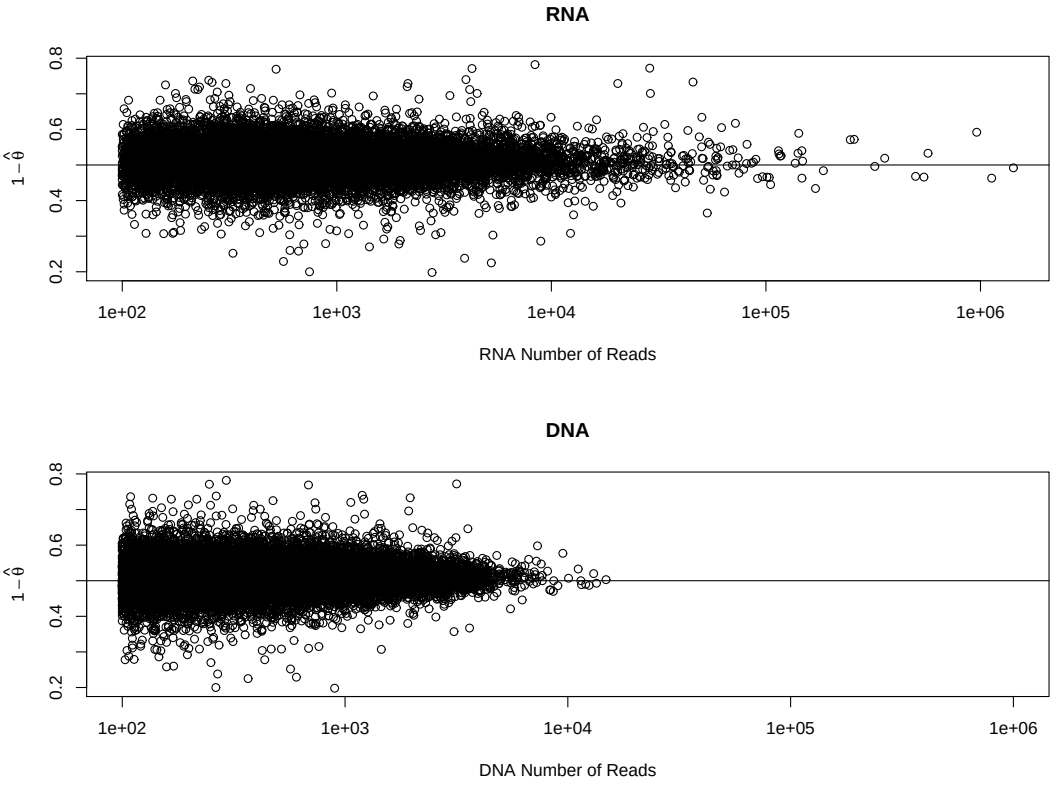


Figure S4: The estimate of AI for RNA $(1 - \hat{\theta})$ is plotted on the y-axis. The x-axis is the number of reads RNA (DNA) across all samples. Data are from exons analyzed with the Bayesian model (n=14,452).

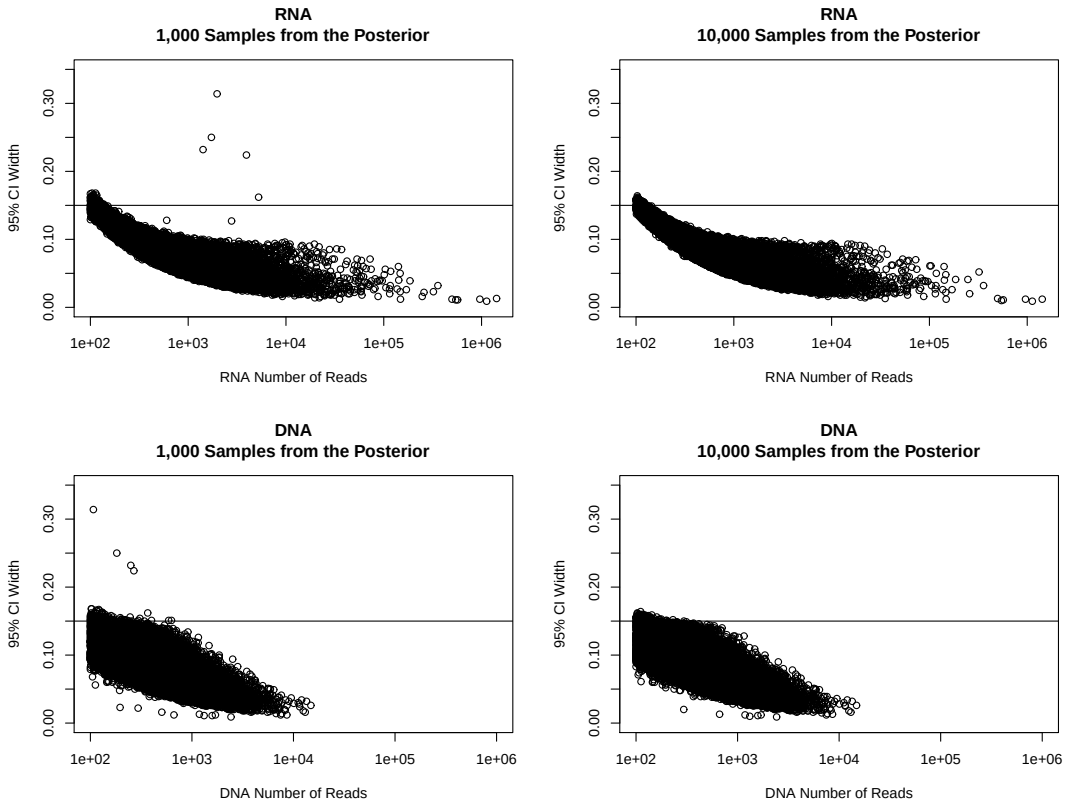


Figure S5: The relationship between credible interval (CI) width and number of reads. The y-axis is the width of the 95% Bayesian credible interval. The x-axis is the number of reads in all samples. Data are from exons analyzed with the Bayesian model (n=14,452). The left panel is for $1,000$ samples from the posterior and the right panel is for $10,000$ samples from the posterior.

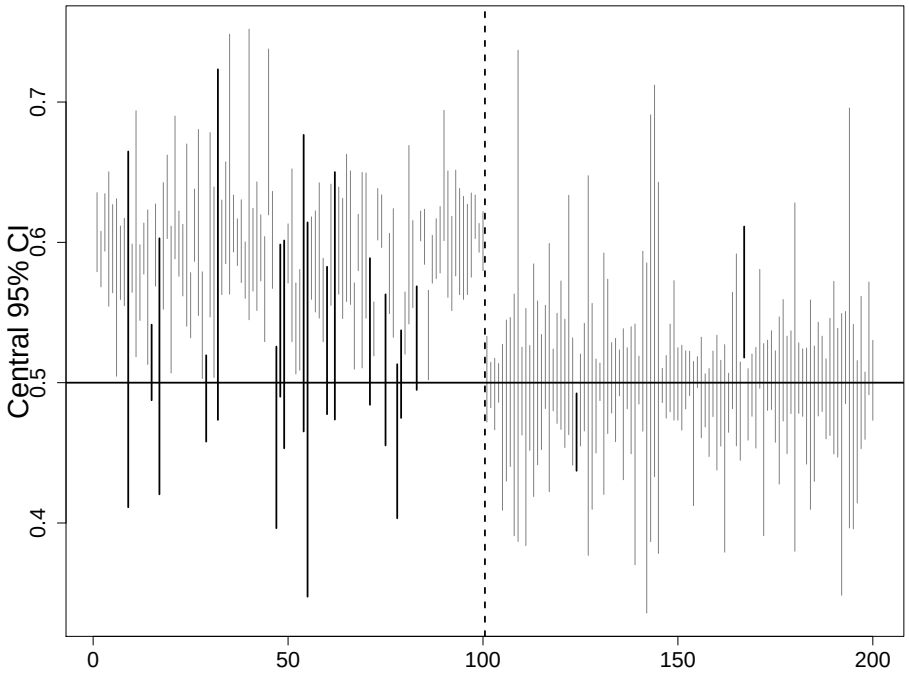
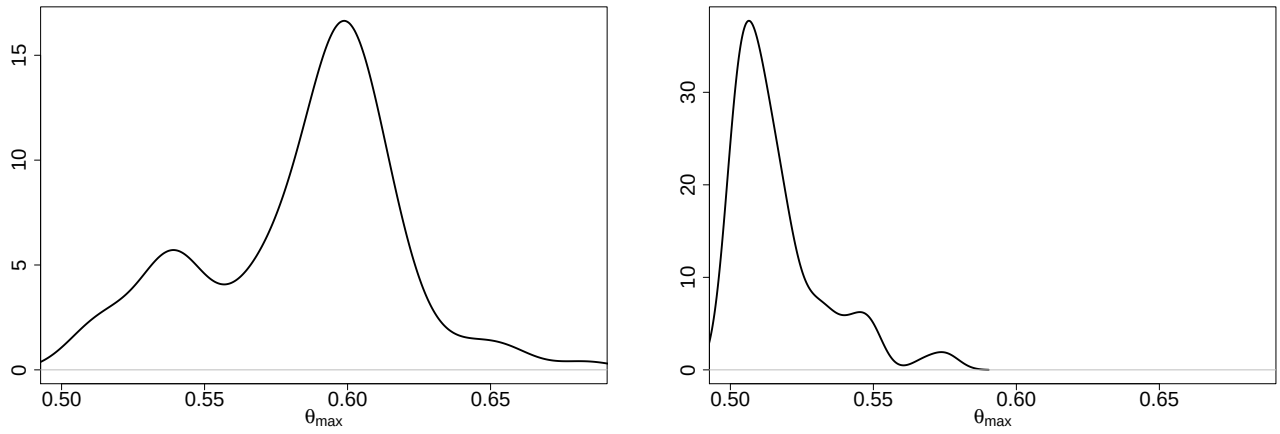
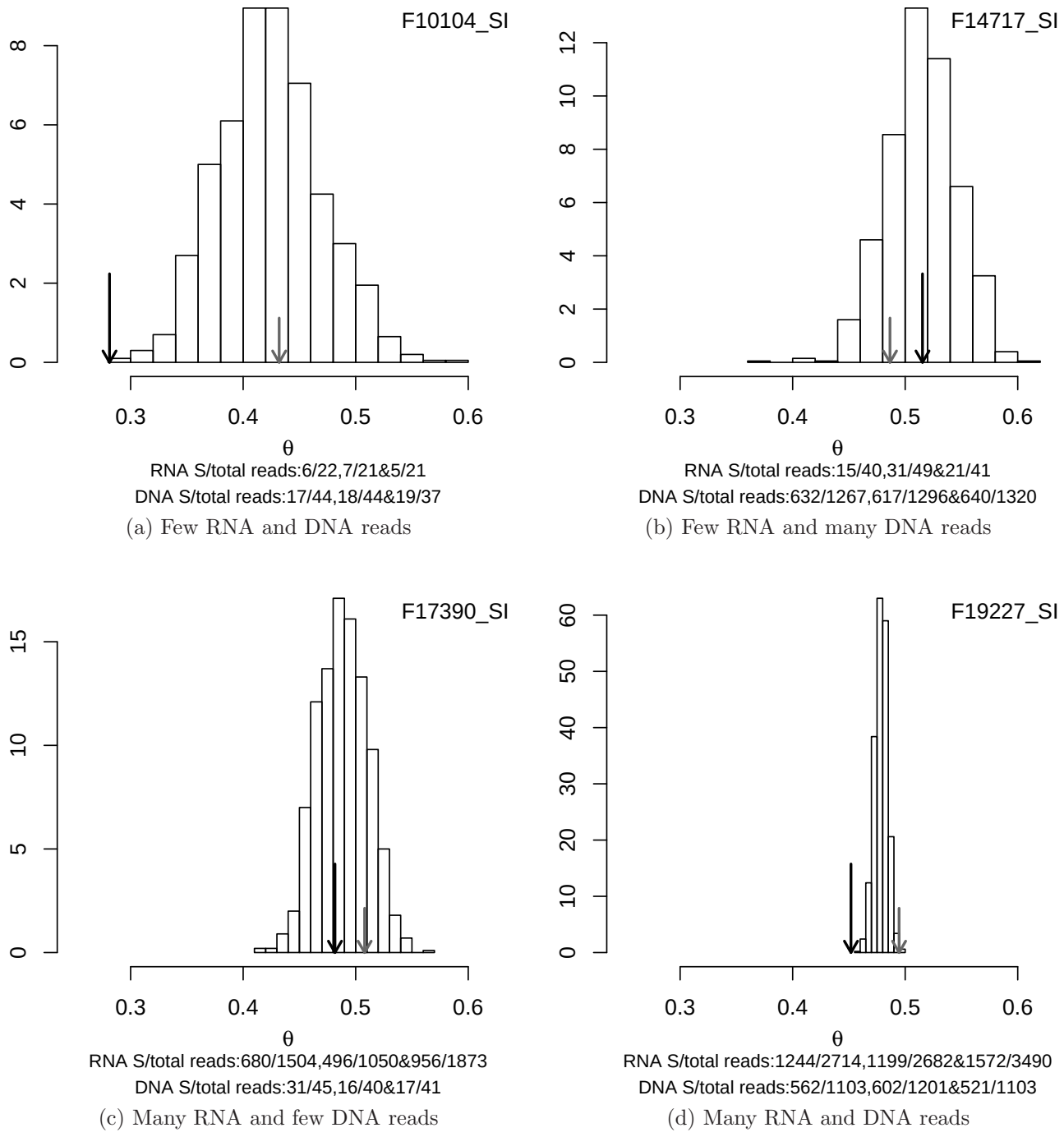


Figure S6: Central 95% Credible intervals (CI) for θ. Only the first 100 simulated gene reads exhibit AI. When this CI does contains 0.5 the gene is flagged as exhibiting AI. Darker CIs to the left (right) of the noncontinuous line represent false negatives (positives). There are 17 false negatives and 2 false positive. For reference we plot the horizontal line through 0.5.



AI No AI

Figure S7: Smoothed histogram for the simulated data for $\theta_{\max} = \max\{\tilde{\theta}, 1-\tilde{\theta}\}$, for exons with AI (right) and without AI (left). Here $\tilde{\theta}$ is the estimate (posterior mean) of θ. The $\theta_{\max}$s for exons with AI are farther from 0.5 than exons without AI.



(a) Few RNA and DNA reads

(b) Few RNA and many DNA reads

(c) Many RNA and few DNA reads

(d) Many RNA and DNA reads

Figure S8: Posterior Distribution of the proportion of S reads in the RNA samples, θ, in four scenarios with different RNA and DNA total reads. The black long (grey short) arrow indicates the recorded proportion of $\#S$ reads in the RNA (DNA) pooled samples. The black long arrow is a natural estimate of the proportion of S reads in the sample if we do not considering DNA information.